Genetics and Molecular Research

HydroCalc Proteome: a tool to identify distinct characteristics of effector proteins

G.J. da Silva[1,2], R.G.T.M. da Silva[1,2], V.A. Silva[1,2], E. C. Caritá[1], A.L. Fachin[1] and M. Marins[1]

[1]Unidade de Biotecnologia, Universidade de Ribeirão Preto, Ribeirão Preto, SP, Brasil
[2]Instituto Federal de Educação Ciência e Tecnologia do Sul de Minas Gerais, Muzambinho, MG, Brasil

Corresponding author: M. Marins
E-mail: mmarins@gmb.bio.br

Genet. Mol. Res. 15 (3): gmr.15038111
Received November 19, 2015
Accepted February 26, 2016
Published July 29, 2016
DOI http://dx.doi.org/10.4238/gmr.15038111

ABSTRACT. Bacterial pathogenicity is associated with secretion of effector proteins into intra- and extracellular spaces. These proteins interfere with cellular processes such as inhibition of phagosome-lysosome fusion, induction of apoptosis and autophagy, activation and suppression of kinases, regulation of receptor activity, and modulation of transcription factors. Knowledge regarding the characteristics of these proteins would assist in pathogenicity studies, and help to identify possible and novel targets for antibacterial drugs. Amino acid hydropathy is a property that can affect behavior patterns in effector proteins. The HydroCalc Proteome tool analyzes total hydropathy, average hydropathy, C-terminal hydropathy, C-terminal load, and basic polar amino acids at the C-terminus. These five properties could contribute to the identification of proteins with an effector potential. HydroCalc Proteome is a web tool that provides a simple interface

for the analysis of hydropathy properties in proteins. This tool permits the analysis of a single protein or even the complete proteome, which cannot be achieved by using other hydropathy tools. The tool displays the result of five properties related to effector proteins in a single table. The HydroCalc Proteome (www.gmb.bio.br/hydrocalc) is a powerful tool for protein analysis, and can contribute to the study of effector proteins.

Key words: Effector protein; Hydropathy; Bioinformatics; Bacterial pathogenicity

INTRODUCTION

Bacterial diseases are associated with the secretion of effector proteins into eukaryotic cells. For example, bacteria belonging to the Anaplasmataceae family, infect various types of mammalian cells, and use type IV secretion systems to secrete effector proteins that modulate cellular processes such as phagosome-lysosome fusion, apoptosis, and autophagy. In addition, they also modulate the activities of kinases, receptors, as well as transcription factors (Alvarez-Martinez and Christie, 2009; Rikihisa et al., 2009). Alterations to these cellular processes induce favorable environments for the proliferation of these obligate intracellular bacteria and their escape from the immune system.

Effector proteins possess structural characteristics that can be identified by using bioinformatic tools. Hydrophilic properties such as total hydropathy, average hydropathy, C-terminal hydropathy, C-terminal load, and number of basic polar amino acids at the C-terminus have been identified in effector proteins of *Bartonella henselae*, *Legionella pneumophila*, *Agrobacterium tumefaciens*, and *Anaplasma marginale* (Lockwood et al., 2011; McDermott et al., 2011).

The average hydropathy of a protein is its total hydropathy divided by the number of amino acids in the protein. Known effector proteins exhibit negative average hydropathy. The proteins identified as effectors contain a hydrophilic C-terminus, in which the translocation signal of the protein is located at the C-terminus (Lockwood et al., 2011). The C-terminal load of proteins is calculated by attributing a load of +1 to basic polar amino acids (HRK) and a load of -1 to acidic polar amino acids (ED). Hence, the overall load is given by the difference between the number of basic and acidic polar amino acids. The positive charge at the C-terminus is an important feature for the identification of effector proteins (Vergunst et al., 2005). The effector proteins identified to date contain at least three basic polar amino acids at the C-terminus (Meyer et al., 2013).

The objective of the present study was to describe the HydroCalc Proteome tool, which can be used for analysis of hydropathy-related properties to help identify and characterize potential effector proteins. This study is justified by the importance of the analyses proposed and by the lack of specific tools that can analyze the entire proteome. Organization of the data obtained with this approach will contribute to understanding of bacteria pathogenicity and identification of antibacterial drugs.

MATERIAL AND METHODS

The bioinformatic tool, HydroCalc Proteome, was developed to analyze hydropathy-

related properties. This tool analyzes five properties: total hydropathy, average hydropathy, C-terminal hydropathy, C-terminal load, and basic polar amino acids at the C-terminus. Calculation of hydropathy is based on the Kyte-Doolittle scale (Kyte and Doolittle, 1982).

This tool was developed using the web technologies HTML 5, CSS 3, and JavaScript (framework Bootstrap), which are able to achieve the desired functions (Goodman et al., 2010). The Apache web server was used to execute web applications.

To use this online tool, the researcher inserts the FASTA file, the hydropathy cut-off, and C-terminus value. The Get function picks up the sequences (in FASTA format), the hydropathy value to be considered for filtering, and the value related to the number of amino acids at the C-terminus. The Vector_sequences function places each protein at the position of a vector (a set), and identify each protein with the character ">" (which precedes each protein in the FASTA file). In this vector, each element possesses its own identification (ID) and its sequence of amino acids. If the vector of sequences is not empty, the current element of the vector is passed to the Format_sequence, Length and Hydro functions. The Format_sequence function formats the sequence, removing blanks and/or tabs. The Length function returns the number of amino acids that the sequence possesses. The Hydro function returns the value referring to the total hydropathy of the protein according to Kyte-Doolittle (1982). If the value returned by the Hydro function is lower than the cut-off entered by the user, the analyses continues. Otherwise, this element is eliminated from the vector. If positive, the sequence is passed to the Avg, Cterm_hydro, Cterm_charge, and Cterm_HRK functions, and a vector of results, Results_vector, is generated. The Avg function returns the average hydropathy per amino acid, i.e., Hydro/Length. The Cterm_hydro returns the C-terminal hydropathy according to Kyte-Doolittle (1982). The Cterm_charge function returns the C-terminal load, attributing a load of +1 to amino acids H, K and R, and a load of -1 to amino acids E and D. Their sum will be calculated as the final load. The Cterm_HRK function returns the number of basic polar amino acids at the C-terminus. Results_vector is the vector of results. The system repeats all previous steps for all proteins. At the end, the Table function returns a table with all calculated results. Figure 1 describes this process in the flow chart format.

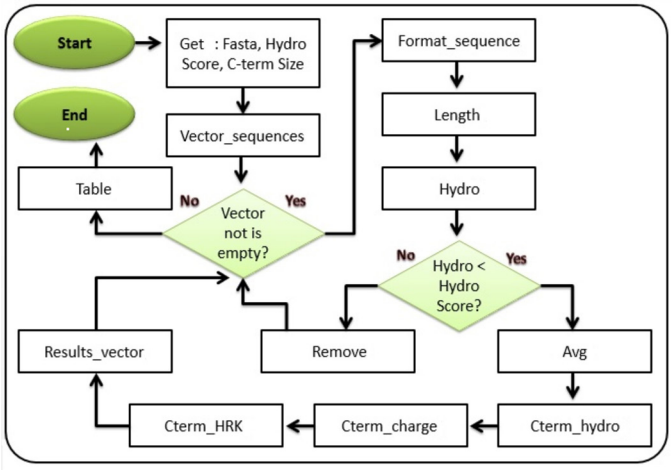

Figure 1. Representative flow chart of the HydroCalc Proteome tool. Logical sequence of the functions of the tool is outlined above.

RESULTS

The HydroCalc Proteome tool has a user-friendly interface, and provides clear information that facilitates its use. The result is an objective interface that is easy to use and handle. It has a window with a scroll bar for entry of the FASTA sequence. The complete proteome can be inserted, and proteins can then be filtered based on hydropathy criteria, selecting hydrophilic or hydrophobic proteins with indices lower than the parameter indicated. It is also possible to choose the number of C-terminal amino acids. Even though it is a web tool, HydroCalc Proteome shows good performance even when more than 1000 proteins are analyzed. The tool is available online, and can be accessed from any platform and executed using any modern internet browser. Figure 2 shows the interface of the HydroCalc Proteome tool.

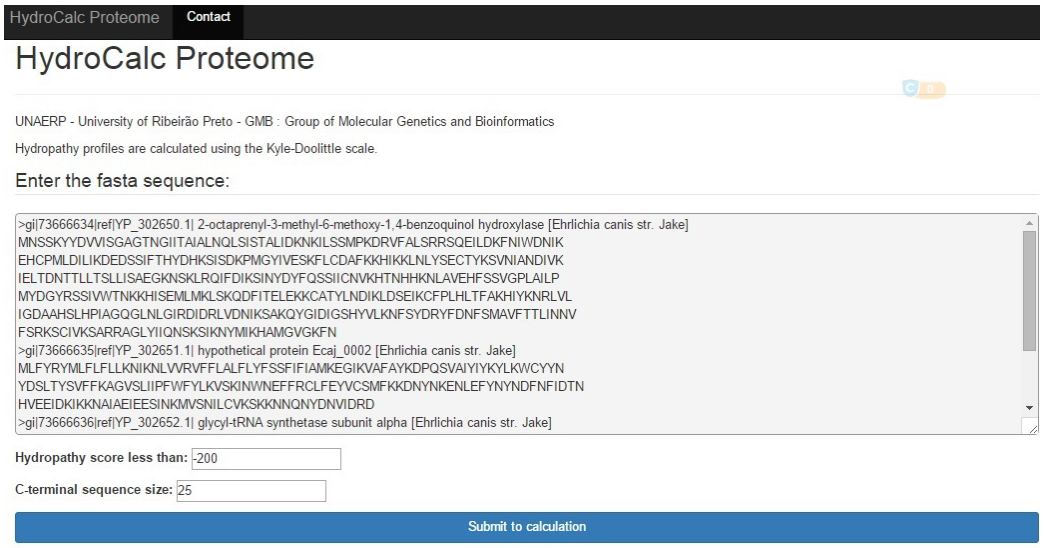

Figure 2. Interface of the HydroCalc Proteome tool. The input data are the FASTA file, hydropathy score, and size of the C-terminus.

The HydroCalc Proteome tool performs the calculations and displays the results in a single table, in which the proteins are shown in rows and their properties are listed in columns (Figure 3). These data are displayed on the screen using the vector of results. The first column contains Nr, with the number of proteins analyzed and selected. The ID column indicates the identification of the protein as inserted into the FASTA file. The Length column shows the number of amino acids of the protein. The Hydro column displays the hydropathy result, with negative values indicating hydrophilic proteins and positive values indicating hydrophobic proteins. The Avg column shows the average hydropathy per amino acid. The C-term hydro column shows the C-terminal hydropathy. The C-term charge column indicates the load of the C-terminus. The C-term HRK column shows the number of each HRK amino acid found at the C-terminus in order.

HydroCalc Proteome | Contact

Result: Hydropathy score less than = -200 and C-term sequence size = 25

Nr	ID	Length	Hydro	Avg	C-term hydro	C-term charge	C-term HRK
1	gi|73666638|ref|YP_302654.1| chaperone protein DnaJ [Ehrlichia canis str. Jake]	382	-253.60	-0.66	-22.90	2	0,1,3
2	gi|73666650|ref|YP_302666.1| gp140 [Ehrlichia canis str. Jake]	688	-604.20	-0.88	1.00	-3	0,0,1
3	gi|73666651|ref|YP_302667.1| type IV secretion system protein VirB8 [Ehrlichia canis str. Jake]	723	-396.40	-0.55	-68.10	-4	1,0,4
4	gi|73666693|ref|YP_302709.1| hypothetical protein Ecaj_0060 [Ehrlichia canis str. Jake]	3714	-1189.70	-0.32	-7.20	1	0,1,1
5	gi|73666695|ref|YP_302711.1| hypothetical protein Ecaj_0062 [Ehrlichia canis str. Jake]	957	-318.60	-0.33	-3.90	3	1,1,2
6	gi|73666696|ref|YP_302712.1| hypothetical protein Ecaj_0063 [Ehrlichia canis str. Jake]	705	-571.50	-0.81	30.70	5	1,1,3
7	gi|73666699|ref|YP_302715.1| hypothetical protein Ecaj_0066 [Ehrlichia canis str. Jake]	889	-487.40	-0.55	-25.30	2	1,1,2
8	gi|73666700|ref|YP_302716.1| hypothetical protein Ecaj_0067 [Ehrlichia canis str. Jake]	695	-232.90	-0.34	-25.80	0	0,3,1
9	gi|73666701|ref|YP_302717.1| hypothetical protein Ecaj_0068 [Ehrlichia canis str. Jake]	616	-230.80	-0.37	-13.50	0	1,3,0

Figure 3. Table generated with the HydroCalc Proteome tool. Each row shows the result of five analyses of the protein.

Following data analysis using the HydroCalc Proteome tool, a group of selected proteins can be generated that the researcher can use to verify distinct properties in effector proteins.

In the present study, four proteins of *A. marginale* that were identified computationally and confirmed experimentally as effector proteins were analyzed. The HydroCalc Proteome tool identified five effector protein characteristics in these proteins. As shown in Figure 4, these proteins were found to be hydrophilic, have a negative average hydropathy, possess a hydrophilic C-terminus and a positive C-terminal load, and contain three basic polar amino acids at the C-terminus.

Result: Hydropathy score less than = -200 and C-term sequence size = 25

Nr	ID	Length	Hydro	Avg	C-term hydro	C-term charge	C-term HRK
1	gi|56387728|gb|AAV86315.1| hypothetical protein AM185 [Anaplasma marginale str. St. Maries]	798	-583.60	-0.73	-35.00	4	0,4,1
2	gi|56387920|gb|AAV86507.1| hypothetical protein AM470 [Anaplasma marginale str. St. Maries]	1261	-804.00	-0.64	-13.30	4	1,3,1
3	gi|56388085|gb|AAV86672.1| ankyrin [Anaplasma marginale str. St. Maries]	1387	-552.80	-0.40	-24.90	3	0,1,2
4	gi|56388400|gb|AAV86987.1| hypothetical protein AM1141 [Anaplasma marginale str. St. Maries]	367	-302.30	-0.82	-52.30	7	2,4,1

Figure 4. Result of effector proteins of *Anaplasma marginale.* Proteins exhibited positive results for the characteristics analyzed with the HydroCalc Proteome tool.

Next, the 145 known effector proteins of *L. pneumophila* were analyzed (Burstein et al., 2009). Results showed that 95.9% of the proteins are hydrophilic, 95.9% have negative average hydropathy, 89.7% possess a hydrophilic C-terminus, 38.6% have a positive C-terminal load, and 73.8% contain more than two basic polar amino acids at the C-terminus.

DISCUSSION

There are existing tools that analyze and identify hydrophobic and hydrophilic regions in proteins based on different scales.

The ProtScale tool, available at http://web.expasy.org/protscale/ (accessed May 10, 2015), generates hydropathy graphs using different scales. The Kyte-Doolittle scale is the standard scale (Wilkins et al., 1999). However, these analyses performed one protein at a time, and the tool does not report other hydropathy properties of effector proteins.

The Platinum tool, available at http://model.nmr.ru/platinum/ (accessed June 10, 2015), analyzes hydrophilic and hydrophobic properties using three-dimensional (3-D) structures. The tool displays a 3-D model of the protein, which permits identification of α-helix conformations (Pyrkov et al., 2009). However, the computational requirements are high, and the tool is not easily applied for analysis of a large number of proteins.

Prediction tools for effector proteins are sparse. The two tools currently used, T4EffPred and S4TE, are limited in capabilities. The T4EffPred tool, available at http://bioinfo.tmmu.edu.cn/T4EffPred/index.html (accessed August 19, 2014), can be used for the prediction of the effector proteins, but the web version only permits the analysis to be carried out with one protein at a time (Zou et al., 2013). Furthermore, this tool was presented with execution errors in terms of access to the database during majority of the attempts made for the analysis. The S4TE tool, available for download at http://sate.cirad.fr/ (accessed April 20, 2015), is a tool for predicting type IV effector proteins (Meyer et al., 2013). However, this tool is not directly available on the web. Technical computational knowledge is necessary for its usage, and it is not available for Windows platforms.

Computational methods are currently used to define a group of proteins for laboratory tests (Lockwood et al., 2011; Wang et al., 2014). Using *A. marginale* and *L. pneumophila* effector proteins as test samples, the HydroCalc Proteome tool was able to identify characteristics in these proteins, which are present in the effector proteins. It is observed that proteins identified as effector proteins possess hydrophilic properties. HydroCalc Proteome is the only web tool that analyzes five properties of effector proteins: total hydropathy, average hydropathy, C-terminal hydropathy, C-terminal load, and basic polar amino acids at the C-terminus. This tool may be useful in screening proteins from the proteome that may be potential effector proteins.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank Instituto Federal de Educação Ciência e Tecnologia do Sul de Minas Gerais for general support and Kerstin Markendorf for English revision of the manuscript.

REFERENCES

Alvarez-Martinez CE and Christie PJ (2009). Biological diversity of prokaryotic type IV secretion systems. *Microbiol. Mol. Biol. Rev.* 73: 775-808. http://dx.doi.org/10.1128/MMBR.00023-09

Burstein D, Zusman T, Degtyar E, Viner R, et al. (2009). Genome-scale identification of Legionella pneumophila effectors using a machine learning approach. *PLoS Pathog.* 5: e1000508. http://dx.doi.org/10.1371/journal.ppat.1000508

Goodman D, Morrison M, Novitski P and Gustaff Rayl T (2010). 7th ed. JavaScript Bible. Wiley Publishing, Inc., Indianapolis.

Kyte J and Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105-132. http://dx.doi.org/10.1016/0022-2836(82)90515-0

Lockwood S, Voth DE, Brayton KA, Beare PA, et al. (2011). Identification of Anaplasma marginale type IV secretion

system effector proteins. *PLoS One* 6: e27724. http://dx.doi.org/10.1371/journal.pone.0027724

McDermott JE, Corrigan A, Peterson E, Oehmen C, et al. (2011). Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infect. Immun.* 79: 23-32. http://dx.doi.org/10.1128/IAI.00537-10

Meyer DF, Noroy C, Moumène A, Raffaele S, et al. (2013). Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. *Nucleic Acids Res.* 41: 9218-9229. http://dx.doi.org/10.1093/nar/gkt718

Pyrkov TV, Chugunov AO, Krylov NA, Nolde DE, et al. (2009). PLATINUM: a web tool for analysis of hydrophobic/hydrophilic organization of biomolecular complexes. *Bioinformatics* 25: 1201-1202. http://dx.doi.org/10.1093/bioinformatics/btp111

Rikihisa Y, Lin M, Niu H and Cheng Z (2009). Type IV secretion system of Anaplasma phagocytophilum and Ehrlichia chaffeensis. *Ann. N. Y. Acad. Sci.* 1166: 106-111. http://dx.doi.org/10.1111/j.1749-6632.2009.04527.x

Vergunst AC, van Lier MC, den Dulk-Ras A, Stüve TA, et al. (2005). Positive charge is an important feature of the C-terminal transport signal of the VirB/D4-translocated proteins of Agrobacterium. *Proc. Natl. Acad. Sci. USA* 102: 832-837. http://dx.doi.org/10.1073/pnas.0406241102

Wang Y, Wei X, Bao H and Liu SL (2014). Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 15: 50. http://dx.doi.org/10.1186/1471-2164-15-50

Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, et al. (1999). Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* 112: 531-552.

Zou L, Nan C and Hu F (2013). Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 29: 3135-3142. http://dx.doi.org/10.1093/bioinformatics/btt554